Title

The fate of Arabidopsis thaliana homeologous CNSs and their motifs in the Paleohexaploid Brassica rapa.

Journal

Genome biology and evolution, 5(4)

ISSN

1759-6653

Authors

Subramaniam, Sabarinath
Wang, Xiaowu
Freeling, Michael
et al.

Publication Date

2013

DOI

10.1093/gbe/evt035

Peer reviewed

The Fate of *Arabidopsis thaliana* Homeologous CNSs and Their Motifs in the Paleohexaploid *Brassica rapa*

Sabarinath Subramaniam[1,*], Xiaowu Wang[2], Michael Freeling[1,*], and J. Chris Pires[3]

[1]Department of Plant and Microbial Biology, University of California, Berkeley

[2]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China

[3]Division of Biological Sciences, University of Missouri

*Corresponding authors: E-mail: freeling@berkeley.edu; shabari@berkeley.edu.

Accepted: March 1, 2013

Abstract

Following polyploidy, duplicate genes are often deleted, and if they are not, then duplicate regulatory regions are sometimes lost. By what mechanism is this loss and what is the chance that such a loss removes function? To explore these questions, we followed individual *Arabidopsis thaliana–A. thaliana* conserved noncoding sequences (CNSs) into the *Brassica* ancestor, through a paleohexaploidy and into *Brassica rapa*. Thus, a single Brassicaceae CNS has six potential orthologous positions in *B. rapa*; a single *Arabidopsis* CNS has three potential homeologous positions. We reasoned that a CNS, if present on a singlet *Brassica* gene, would be unlikely to lose function compared with a more redundant CNS, and this is the case. Redundant CNSs go nondetectable often. Using this logic, each mechanism of CNS loss was assigned a metric of functionality. By definition, proved deletions do not function as sequence. Our results indicated that CNSs that go nondetectable by base substitution or large insertion are almost certainly still functional (redundancy does not matter much to their detectability frequency), whereas those lost by inferred deletion or indels are approximately 75% likely to be nonfunctional. Overall, an average nondetectable, once-redundant CNS more than 30 bp in length has a 72% chance of being nonfunctional, and that makes sense because 97% of them sort to a molecular mechanism with "deletion" in its description, but base substitutions do cause loss. Similarly, proved-functional G-boxes go undetectable by deletion 82% of the time. Fractionation mutagenesis is a procedure that uses polyploidy as a mutagenic agent to genetically alter RNA expression profiles, and then to construct testable hypotheses as to the function of the lost regulatory site. We show fractionation mutagenesis to be a "deletion machine" in the *Brassica* lineage.

Key words: conserved noncoding sequence, CNS, fractionation, mutagenesis, deletion, G-box, *PIL5*, *Arabidopsis*, *Brassica rapa*.

Introduction

A perplexing and long-standing problem in classical genetics is to know when a recessive mutant specifies a complete knockout of function. Even the sequence of mutants with recessive phenotypes compared with the wild type progenitor may not answer the question of functionality. If the mutation happened during evolution, inferred from comparisons of mutant with a more ancestral outgroup, it is even more difficult to predict functionality. Because of the history of paleopolyploidy in all plant lineages (Van de Peer 2011), updated at (http://synteny.cnr.berkeley.edu/wiki/index.php/Sequenced_plant_genomes, last accessed March 23, 2013), and the consequent potential for functional redundancy, duplicate genes, or regulatory site sequences mutate into nondetectability. This postpolyploidy gene loss, called fractionation, is widespread and frequent. It is important to know whether such "loss" results in loss of function. One way to show that a loss of sequence detectability is a loss-of-function is to show that the loss is by deletion of sequence, because a deleted sequence cannot function.

When a genome doubles or triples, as with paleotetraploids or hexaploids, each chromosome with each gene is initially duplicated. What follows is a process of chromosomal evolution called "diploidization," during which the polyploid becomes rearranged and altered to act as a meiotic diploid (Wolfe 2001). The newly diploidized polyploid tends to lose one or the other of its duplicate genes (fractionation), usually much of the time, as expected in theory (Lynch and Force 2000) and realized in practice (Sankoff et al. 2010). The fractionation mechanism is a sort of intrachromosomal

recombination inferred from short repeats flanking progenitor-deleted sequences (Petrov et al. 1996; Devos et al. 2002), and is known for postpaleotetraploid maize (Woodhouse et al. 2010) and postpaleohexaploid *Brassica rapa* (*Br*) (Tang et al. 2012). Even if a gene pair survives polyploidy, perhaps because of subfunctionalization (Lynch and Force 2000) or tendency to maintain product dosage balance (Freeling 2009), that does not mean that all parts of the gene will remain duplicated. This study follows individual conserved noncoding sequences (CNSs) known to exist around many *Arabidopsis* genes as they now exist in *Br*, a hexaploid. Figure 1 follows one ancestral Brassicaceae gene as it gets duplicated during the alpha paleotetraploidy, and then follows as each alpha homeolog splits into the lineage that will be *Arabidopsis* (*At*) or *Br*, and then through the *Brassica* lineage, on through the paleohexaploidy, and finally follows the genes into the six potential chromosomal positions on the three *Br* subgenomes. Sometimes a *Br* gene is fractionated and takes all of its *At*-orthologous CNSs with it, but sometimes the duplicate transcriptional unit and its *cis* sequences persist. In such cases, sometimes the *At* CNS being followed goes undetectable but the gene remains and is transcriptionally active. This has been shown previously in grasses (Schnable et al. 2011). The red arrow on figure 1 denotes such a CNS loss. The small squares decorating the gene models of figure 1 are CNSs.

The mechanism of CNS fractionation in plants has not been studied previously, although it is known that plant CNSs lose detectability as divergence time increases (Reineke et al. 2011). This mechanism is important because several CNSs have been shown to function as cis-acting regulators and are enriched in known DNA-binding motifs (Freeling and Subramaniam 2009; Raatz et al. 2011), they are associated with DNaseI open chromatin (Zhang et al. 2012) and with the suppression of gene expression (Spangler et al. 2011). Thus, CNS loss of detectability could predict loss of a specific regulatory function, but only in the case that the CNS loss marks loss of CNS function.

Crucifer CNSs in *Arabidopsis* have a history. Previous work (Thomas et al. 2007) found that 14,944 CNSs (alpha-CNSs, αCNSs, *At–At* CNSs) retained following the most recent tetraploidy in the *Arabidopsis thaliana* (*Arabidopsis*, *At*) lineage. Genes retained as pairs following this tetraploidy, called homeologs (or homoeologs, Ohnologs, syntenic paralogs), have diverged a modal 0.76 in synonymous base substitution rate (Ks), and this was shown to be an adequate evolutionary divergence proxy to ensure that associated CNSs avoided purifying selection because CNSs on average, functioned. When divergence times become too great (>0.9 modal Ks), detection of CNSs becomes difficult, and when there is too little divergence, or when the detection algorithm settings are set without regard to noise levels (Kaplinsky et al. 2002; Thomas et al. 2007; Lyons and Freeling 2008), CNSs no longer indicate putative conserved function.

We know enough about the genome of *Br* to make some predictions. The three ancestral genomes of the new *Br* hexaploid do not remain intact for long. Fractionation soon removed most of the redundant duplicated genes (Wang et al. 2011) and is predicted to have removed some

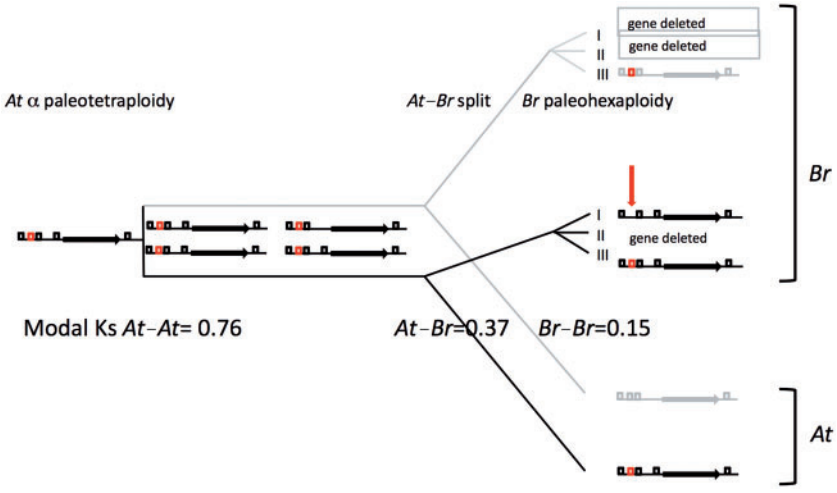

Fig. 1.—The gene tree of a pre-alpha tetraploidy Brassicaceae gene with an protein-coding sequence (black arrow) and five CNSs (boxes on the model line) as it duplicates at the alpha and, in the *Brassica* lineage, undergoes an additional paleohexaploidy before it was sequenced in *Brassica rapa* (*Br*). The modal Ks values, for each of these three events are shown, as downloaded from the SynMap application in the CoGe toolbox. The red CNS exemplifies the sort of CNS we follow. It is detected as a conserved sequence between the two homeologous genomes of *Arabidopsis* (*Arabidopsis thaliana*, *At*), but is fractionated (red arrow) from one of the *Br* homeologs in this doublet. The gray lineage is of the "out-paralog" Brassica lineage, in this case represented in *Br* as a singleton gene. Note that a CNS was lost just 5′ of coding sequence, and is not present in the out-paralog lineage. Even though this CNS did exist in the test lineage, we did not test for it because we began with homeologous *At–At* CNSs.

duplicate CNSs as well. Thus, each CNS in *Br* is retained as a singlet, a doublet, or a triplet depending on whether its gene is retained, and if its gene is retained, depending on whether the CNS itself remains detectable. Some go undetectable, as with the CNS position at the tip of the red arrow of figure 1. The background "neutral" base substitution rate between *Arabidopsis* and *Br* orthologs (0.38, legend fig. 1) will tend, in theory, to substitute nucleotides in CNSs that contribute little or nothing to CNS function. The αCNSs of *Arabidopsis* contain sequences that come with varying blastn *E* values and lengths down to 15 bp. As plant CNSs contain DNA-binding motifs (Freeling and Subramaniam 2009) as they do in mammals (Pennacchio et al. 2007; von Rohr et al. 2007), motifs known to be short and inexact (7–12 bp with alternatives), some of the *At* αCNSs should not be detected in *Br* even though they might contain functional motifs because the motifs are shorter than the minimal length of detectable CNSs. In short, if base substitution were the prevailing mechanism of going nondetectable, then CNSs could drift into nondetectability and still conserve typical, functional DNA-binding motifs. However, no matter what the mechanism of nondetectability, selection for functional loss should be greater for a CNS on a singleton gene, the CNS being more unique-sequence, as compared with CNS on each of a doublet or triplet *Br* gene, the more redundant situation. This inference is the basis of our essential strategy; see the second footnote of table 1 where our strategy is applied to real data.

If a CNS is undetectable using our standard blastn criteria, we use computational methods to deduce the preponderant mechanism of each CNS's mutation from the ancestral sequence, detailed in the Materials and Methods section. As much is known about particular G-boxes within CNSs (Freeling et al. 2007), based on previous work on this motif and transcription factors that bind some G-boxes, we study how G-boxes become nondetectable as well. Our categories of loss: 1) base substitutions (the pseudogene pathway), 2) proved deletions (removal of one or both flanking markers as well as the CNS), 3) computationally inferred deletions, 4) indels, or 5) large insertions. Indels have been recently identified as a significant mutational endpoint in plants (Hollister et al. 2010).

Understanding how plant CNSs go undetectable is important for several reasons. In animals, explanations have been proposed for how undetectable enhancer sequences sometimes retain function, including "binding site turnover" (Hancock et al. 1999; Ludwig et al. 2000; Dermitzakis and Clark 2002) and "dormant TF-binding sites" (Junion et al. 2012). These mechanisms require nonfunctional sequences drifting along the pseudogene pathway before they mutate back to function. Such mechanisms become less likely as nonfunctional DNA is deleted more quickly in plants. Additionally, knowing the mechanism of CNS fractionation is particularly important in light of a genetic-type method we have proposed called "fractionation mutagenesis" (Freeling et al. 2012). This

Table 1
Gene Redundancy Was Used to Infer Whether an Undetectable *At* CNS Is Undetectable in *Br* Because It Has Been Mutated into Nonfunction and Categorization by "Predominant Cause" Was Done Computationally

Predominant Cause of *At–At* CNS Loss in *Br* Determined by Computation	***B/A***	**Total No. of Undetectable 5′ + 3′ CNSs > 30 Bases[a]/Total (%)**		
	Functionality Metric[b] **1 = functional**	***A* = *N* (%) for CNSs in Singlet *Br* Genes (Control)**	***B* = *N* (%) for CNSs in Doublet or Triplet *Br* Genes**	***A* + *B***
		—	—	—
1. **Segmental deletions**, flanking feature(s) lost. Functionless.	5.3	16/966 (~1.6%)	129/1,543 (8.4%)	145/2,509
2. **Computed deletions**, but with no flanking feature also deleted	3.4	43/966 (4.45%)	226/1,543 (15.2%)	269/2,509
3. **Indels**, both insertions and deletions of any size.	2.0	22/966 (2.2%)	90/1,543 (4.4%; 2.3% in doublets, 8.9% in triplets)	112/2,509
4. **Insertions**, of any length	1.1	14/966 (1.4%)	37/1,543 (1.5%)	51/2,509
5. **Base pair substitutions**	1.8	11/966 (1.13%)	16/1,543 (2%, 1.5% in doublets; 4.14% in triplets)	27/2,509
All *At* CNSs undetectable in *Br*	3.8	106/966 (10%)	498/1,543 (33%)	604/2,509
All *At* CNSs detectable in *Br*	Nonsense	860 (89%)	1,045 (67%)	1,905/2,509

[a]Almost all CNSs in *Arabidopsis* that are >30-bp long cannot be rendered blastn-undetectable by the *At–Br* modal Ks of 38% (fig. 1).
[b]Our logic: we assume that a *Br* gene that has been fractionated down to a singlet will tend to require the function of all of its ancestral CNSs, but that CNSs on doublet or triplet genes will be subject to occasional loss of function due to redundancy. Thus, any predominant cause category that tends to NOT reduce function should have *B/A* ratio nearer to 1. Thus, *B/A*, the functionality metric, varies from 1 (functional) to 5.3 (functionless).

method quantifies the RNA levels of duplicate genes in a polyploid and also compares each homeolog's CNS-loss pattern. A mutant expression pattern is then associated with a lost CNS or a cluster of CNSs, and these previously mysterious sequences acquire a testable hypothesis as to ancestral function. We will show that this method of fractionation mutagenesis comprises a natural "deletion machine" 84% of the time in the posthexaploidy *Brassica* lineage and deletion mutations are certainly loss-of-function.

Materials and Methods

Rationale for Confining This Article to CNSs Defined in One Specific Way

Our *At–At* CNSs reflect one definition of a CNS: a syntenic noncoding conservation detected by blastn with significance at or better than a 15/15 exact match, and between genomes or subgenomes diverged to a modal Ks of 0.9–0.5 (Freeling and Subramaniam 2009). The CNS data set produced has the advantages of having been the object of some study, and because this data set depends on local alignments generating an even-handed sampling of conserved noncoding regions no matter how far they may exist from any conserved coding sequence. It is certainly more sensitive to anchor on a coding part of the gene and extend alignments, but this sensitivity only applies close to the anchor. For example, multiple global alignments anchored on the start of transcription and moving up to 1 kb 5′ have provided an excellent CNS data set, and they do overlap with ours (Baxter et al. 2012), but this data set goes deficient in those CNSs moving from 500 bp to 15 kb away from the nearest exon; transposon insertions disrupt global alignments. No one method of obtaining CNSs is best. Fortunately, complete coverage is not important for us to see how *Arabidopsis* CNSs are lost in the posthexaploidy *Br* lineage (fig. 1), so we use homeologous *Arabidopsis* CNSs from our *At–At* v2 data set (supplementary table S1, Supplementary Material online) described later.

Arabidopsis αCNSs

In 2006, there were no sequenced Brassicaceae genomes within the window of Ks 0.5–0.9. However, the two alpha *A. thaliana* subgenomes descended from its most recent paleotetraploidy were nicely diverged for CNS discovery, so the TAIR4 version of the *Arabidopsis* genome was compared with itself (Thomas et al. 2007). Of the original 14,944 individual *At–At* CNS sequences of version 1, 3,635 CNSs were removed: 82 were found to be out of synteny, 22 erroneous CNS calls, 169 CNSs were reassigned to neighboring genes, 1,831 CNSs were invalidated due to wrong direction, and 1,531 CNSs were found to overlap coding DNA sequences (CDSs) or RNA genes, called as annotation of plant genomes became more complete. Version 2 CNSs, identified in relation to TAIR8 annotations, are syntenous in relation to other homeologous features. Column *A* of supplementary table S1, Supplementary Material online, is a notation for each of these version 2 CNSs that includes the *At* gene name to which each sorts; the actual sequence of this sequence is displayed later in the row. Our CNS calls may be proofed easily with the GEvo links of supplementary table S1, Supplementary Material online. GEvo is the sequence comparison tool in the CoGe toolbox (http://genomevolution.org, last accessed March 23, 2013) of comparative genomics applications (Lyons and Freeling 2008). GEvo provides a graphical comparison of multiple genomic regions indicating high-scoring segment pairs (HSPs) for a variety of sequence similarity testing algorithms (selected under the "Algorithm" tab of GEvo), between the defined genomic regions. GEvo was used extensively during our version 1 to version 2 update. The 11,302 version 2 *At–At* CNSs have been "burnt" onto a TAIR8 genome on the model line—this genome is identified as id 39598 in CoGe. These CNSs can be visualized within GEvo by selecting "Yes" for "Show pre-annotated CNSs" under the "Results Visualization Options" tab in GEvo. Using GEvo, our precalled CNS positions can be readily compared through HSPs generated by blastn (default blastn settings with a spike of 15 nucleotides). Column *B* of supplementary table S1, Supplementary Material online, is easy to parse for CNS length; we focus only on those 2,509 longer CNSs for our focal experiment in which we define a functionality metric for each category of αCNS loss (table 1). However, all version 2 CNSs are used for other experiments and all motif experiments.

Locating Orthologous Coordinates for αCNSs within *Br*

For each of the 6,330 *At* genes with a retained *At* α-homeolog (supplementary table S1, Supplementary Material online), we used the synteny screening blocks technique (Tang et al. 2011) to identify all possible orthologous regions in the *Br* Chiifu (Chinese cabbage) genome. Given the recent hexaploidy in the *Br* lineage (fig. 1), we expected to find up to three orthologous copies for each *At* gene. Of the 6,330 *At* genes used for version 2 CNS discovery, we (Tang et al. 2012) identified at least one *Br* ortholog for 6,245 of them, with 2,391 *At* genes having a single detectable orthologous copy in *Br* (singlets), 1,723 *At* genes with two orthologous copies in *Br* (doublets), and 654 *At* genes with three orthologous copies in *Br* (triplets).

In the absence of CNS fractionation, and assuming that our CNSs were sorted to their correct gene, we expected to find the CNS whenever we found the gene in *Br*. Based on *At–Br* orthologies, we expected to find 9,179 CNSs within expected orthologous positions, 3,882 as singlets, 3,678 as doublets, and 1,619 as triplets. We attempted to detect each of these 9,179 αCNSs within each expected orthologous *Br* region. Nucleotide sequence of the gene space (expanded genomic region around and including coding regions and spanning from farthest upstream to downstream CNS) of each *At*

gene containing one or more of these 9,179 CNSs was masked for very repeated sequences (50× copies across entire *At* genome). The corresponding gene space of each detected orthologous region (singlet, doublet, or triplet ortholog) in *Br* was also repeat masked. Each *At* and *Br* orthologous genespace pair was compared using the same blast settings used for *At–At* CNS discovery (Thomas et al. 2007). Every blast HSP hit to the *Br* orthologous genespace was then screened for synteny, using a perl script, to filter out probable noise, whereas simple sequences were filtered out using the DUST filter option of BLAST.

αCNSs that do not show a hit using the above blastn settings are valuable data. In the following section, we describe analysis of such sequences using a global alignment algorithm to determine the nature of evolutionary modifications that may have contributed to the lack of detectability of these CNSs. The general idea is this: The CNSs "lost" in a singlet are assumed to still function, but to have drifted in functionless sequence, or to have suffered "binding site turnover" (Moses et al. 2006). Our Discussion section argues that this assumption is not the whole story, but we did make this assumption. Any frequency of nondetectability above the baseline of loss in singletons was interpreted as being caused by actual functional loss either by base substitution, deletion, small indels, or a large insertion.

Identifying the Molecular Mechanisms That Caused the Lack of Detectability of αCNSs in *Br* Doublets and Triplets

Locating Orthologous Coordinates within Br *for αCNSs Undetectable Using Blastn*

Earlier, we described the use of our previously published CNS discovery blastn settings to measure detectability of αCNSs within expected orthologous gene spaces within each of the three *Br* subgenomes. Each αCNS that was not detected in the expected orthologous region of *Br* using our standard settings was retested to determine the predominant mechanisms that could potentially contribute to the lack of detectability. We started by identifying and extracting the nucleotide sequence for the expected orthologous regions for each αCNS that was not detectable within *Br*. As in the case of the blastn analysis used for measuring CNS detectability, we used the *Br* orthologous gene spaces (coding region + 40 kb on both sides of coding region) as the subject sequence. The query sequences were the coding regions of the *At* gene to which the αCNS was assigned in version 2 (supplementary table S1, Supplementary Material online) to which was added the nucleotide sequence extending out to and including the farthest upstream and downstream CNS. In supplementary figure S1, Supplementary Material online, the GEvo link points to a graphic where this *At* gene space is highlighted yellow; to see this requires selecting "see genespace" in the GEvo options panel.

Each pair of *At* and *Br* gene spaces were compared using blastz with default settings. The position of each αCNS was studied for overlap with any blastz HSP (high scoring segment pair) between the *At* and *Br* genespaces. Those CNSs that overlapped with blastz HSPs were assigned the location of the HSP (start and stop positions) as its expected location in *Br*. CNSs that did not overlap a blastz HSP, but were found flanked by blastz HSPs, were assigned an expected position between the flanking HSPs. In cases where flanking HSPs were not present, depending on the position of the CNS relative to the gene, the expected location was defined from either the start of the orthologous genespace to the start position of the gene, or the stop position of the *Br* gene to the end of the *Br* genespace (Tang et al. 2012). An example of this procedure follows. Supplementary figure S1, Supplementary Material online (http://genomevolution.org/r/4dc3, last accessed March 23, 2013) shows an annotated view of the same GEvo panel described earlier (fig. 3), but now displays blastz HSPs between the *At* gene and its *Br* orthologs. In this figure, αCNS 315; 2;CNS_s680 (pink highlight in fig. 3) overlaps with a blastz HSP in *Br* II and *Br* III but falls between 2 flanking HSPs in *Br* I. The search sequences used for studying the mechanisms acting on this CNS in *Br* II and *Br* III are the blastz HSPs labeled *Br* II and *Br* III (supplementary fig. S1, Supplementary Material online). The orthologous region in *Br* I falls between these two blastz HSPs, indicated in pink highlight in supplementary figure S1, Supplementary Material online.

Identifying the Molecular Mechanisms That Caused the Lack of Detectability of αCNSs in Br *Doublets and Triplets*

There are several possible reasons for a CNSs to go undetectable: deletions of an entire chromosomal segment resulting in the removal of one or more CNSs, or relatively smaller scale changes including insertions, smaller deletions, a combination of both (indels) and base substitutions making individual CNSs undetectable. For smaller CNSs, even one base substitution would either destroy the minimum exact match blast wordsize or drop the CNS below the *E* value cutoff, that equal to a 15/15 exact match. We wrote a simple perl script to use a global alignment algorithm (Needleman and Wunsch 1970) with cost-free ends (BLOSUM 62) to align the nucleotide sequences of each αCNS without a detectable ortholog in *Br* with the expected orthologous regions within *Br*; these regions were found as described previously.

Using the genomic positions for the expected location of each CNS in *Br* to inform our search, we generated an alignment between each CNS sequence and the nucleotide sequence (repeat masked) corresponding to the expected location in *Br*. The *Br* and *At* sequences were aligned using a global alignment algorithm (not a blast family algorithm) with no end gap penalties (Needleman and Wunsch 1970). We did this because the CNS sequence length is much shorter than the *Br* subject sequence. A score value was generated for each alignment and a *P* value statistic was used to measure

quality of alignment. As control for the alignment for each αCNS, we used a perl script to generate 10,000 "scrambled" random sequences, each representing a "random" permutation of the nucleotides that make up each CNS. The alignment score for each CNS was compared with those of the 10,000 random sequences to generate the *P* value of significance. We define any alignment with *P* value less than or equal to 0.05 as being "above noise." Each high-quality, optimal alignment generated by the global-npe algorithm was analyzed using perl scripts for deletions, base substitutions, insertions, and exact matches at each position. Alignments with gaps only on the *Br* sequence were classified (using a perl script) as deletions, those with gaps only on the *At* sequence were classified as insertions and when gaps occurred on both *Br* and *At* sequences, such alignments were classified as indels. We then used these "gaps" data to infer the predominant mechanism contributing to lack of detectability of *At–At* α-CNSs in *Br*.

mRNA Levels for *Br* Genes in Seedling Root and Shoot

As our *Br* sequence is from Chiifu, a Chinese cabbage variety, it is important to know that our RNA expression data are from this same genotype. RPKM (reads per kilobase per million mapped reads) data for genes expressed in seedling stem, leaves, and roots has been analyzed and presented in supplementary table S3, Supplementary Material online, (Cheng et al. 2012) as a control experiment for potential gene death.

Revised CNS-Enriched Transcription Factor-Binding Sites Motif List

Using version 1 of the *At* αCNS list, previous work (Freeling et al. 2007) identified a few known transcription factor-binding sites (TFBS), as regular expression motifs, that were significantly enriched in αCNS sequence as compared with noncoding, nonconserved sequence. While the G-box, a "strictly conserved" palindromic hexamer, was by far the most significantly enriched, other "strictly conserved" motifs were significantly enriched over 2-fold as well. By "strictly conserved," we mean that at least 5 nucleotides within the consensus sequence for the motif must be conserved in the same order; e.g., For the G-box, the consensus motif is CACG TG, the core of the consensus motif "ACGTG" should be conserved to be considered "strictly conserved." Because this work begins with a revised CNS list, version 2, and because we wanted to refine how we controlled for nonfunctional motifs (e.g., we did not mask transposons in our previous work), we updated our enriched motif list (supplementary table S4, Supplementary Material online). We did not use all of the often overlapping motifs available, and in the literature, but concentrated only on 12 motifs picked that were, like the G-box, more strictly conserved and enriched by more than 2× in CNSs: C*ACGT*G (the G-Box), 5′*ACGT*GGC (in the *ACGT* category), GCCGCC (jasmonic acid box), 5′AAACCCTA, and 5′CCGTCC (Freeling et al. 2007) to which we added [CT]*ACGT*GGC, C*ACGT*GGC, *ACGT*GGCA, *ACGT*GTC, AAA CCCTAA, TGTCTC, CC*ACGT*GG. Several of these motifs can be seen (italicized) to be similar. This strictly conserved criterion was used, so that we could more easily infer whether they were intact following mutation to nondetectability. Specific references for each motif sequence are in supplementary table S4, Supplementary Material online, and in a beta-test application in CoGe: MotifView (http://genomevolution.org/CoGe/MotifView.pl, last accessed March 23, 2013). Noncoding, nonconserved, and nontransposon regions from *within the same gene space* as each αCNS were used as the control for each of these motif enrichment studies.

αCNSs That Are Reinforced by Overlap with Published Pil3-like5 Protein (PIL5) Binding Sites and Their G-Boxes

Oh et al. (2009) used ChIP-chip (chromatin immunoprecipitation with microarray sequence recognition methods) data to infer that 748 *Arabidopsis* genomic-binding sites were occupied by basic helix-loop-helix transcription factor PIL5 and 166 nearby genes were upregulated directly by PIL5. As PIL5 has been shown to bind CACGTG, each G-box within a PIL5 "peak" represents a strong argument for a functional G-Box. We compared these PIL5 sites for overlap with our αCNSs. These 32 G-boxes were assumed to be particularly likely to be functional. Five of these CNSs did not have an ortholog in *Br*, understanding these is outside of our topic. The remaining 27 were studied at all orthologous positions in *Br*.

The relatively low number of αCNS-PIL5 peak overlaps was expected. αCNSs (not being orthologous CNSs) can only include those cis-acting sites that were retained after the most recent tetraploidy in the *Arabidopsis* lineage. Further, our unanchored blastn pairwise CNS discovery tool, while necessary to find CNSs that are far from coding sequence syntenic anchors, is known to miss many if not most of the cis-acting sites that are close to the transcription unit (Thomas et al. 2007).

Results

At αCNSs Updated to Version 2

The updated version 2 αCNSs list (supplementary table S1, Supplementary Material online) now contains 11,448 sequences or 5,724 α pairs (α pairs are homeologous pairs derived from the most recent whole genome duplication event in the lineage of *A. thaliana*). These αCNSs were used to search for retention within *Br* at orthologous loci. As a control for our CNSs discovered through manual comparison of homeologous regions in *At (v2)*, we ran our automated CNS Discovery Pipeline v3.0 (https://github.com/gturco/find_cns/tree/master/pipeline, last accessed March 23, 2013) over our homeologous gene pairs to generate an automated *At–At* αCNS data set (supplementary table S2, Supplementary Material online). There is 80% concordance between the automated and manually generated CNS data sets; the v2 data

set was used in this study. The αCNSs, (both v2 and the pipeline 3.0, for comparison) have been added to the gene models of TAIR8 in CoGe as genome data set ID = 39598 (http://genomevolution.org/r/4iaq, last accessed March 23, 2013). Our supplementary table S1, Supplementary Material online, includes links to GEvo in CoGe using these customized gene-space models, thus facilitating reproduction and proofing of our results. Figure 2 (http://genomevolution.org/r/4db1, last accessed March 23, 2013) shows GEvo blastn output graphic, where the query is *Arabidopsis At1G75520*, a bigfoot gene encoding a RING zinc finger protein of unknown function, displayed with its corresponding α-homeolog. Both manually curated (v2, color-coded purple) and automated pipeline 3.0-generated αCNSs (color-coded green) are annotated on this graphic along with the blastn HSPs (color-coded orange) corresponding to regions of high sequence similarity between the homeologs.

General Features of Detectability of *At–At* CNSs in *Br*

The paleohexaploidy in the *Br* lineage generated three subgenomes, with one of them (subgenome III) having almost twice as many genes as either of the other two. Genome dominance and purifying selection explain this phenomenon, using the exact same argument that was proven valid in maize. In terms of CNS detectability, we expect the dominant subgenome (III) to carry most of the genes that are singlets, and subgenomes I and II to have endured the most gene and CNS loss.

For each of the 6,330 *At* genes with a retained *At* α-homeolog, each used for CNS discovery, we used our synteny screening blocks technique (Tang et al. 2012) to identify all possible orthologous regions in the *Br* genome. Given that the recent hexaploidy in *Br* occurred following divergence from the *Arabidopsis* lineage (fig. 1), we expected to find up to three orthologous copies for each *At* gene. Figure 3 is a GEvo graphic (http://genomevolution.org/r/4db6, last accessed March 23, 2013) of the same bigfoot gene shown in figure 2, this time showing blastn hits to the three detected orthologous regions within *Br*. The top panel shows the αCNS-rich *At* gene (*AT1G75520*), a member of SHI ring Zn-finger gene family, and the three panels below show its detectable orthologs in *Br*. αCNSs (v2, purple bars) and the gene space (yellow background) are annotated on the *At* gene panel. HSPs between the *At* gene and each *Br* subgenome ortholog is annotated as *Br* I, *Br* II, and *Br* III (fig. 3). Analysis of the overlap of αCNS positions with corresponding HSPs to each of the *Br* orthologous positions in figure 3 gives insight into the detectability of αCNSs in each of the three *Br* regions. One of the *At–At* CNSs (315; 5;CNS_s677) shown in figure 3—highlighted in gray—has corresponding HSPs in *Br* II and *Br* III subgenomes, but has an undetectable ortholog in *Br* subgenome I. Another CNS, highlighted in pink (315; 8;CNS_s680) has corresponding HSPs only in *Br* II subgenome and has undetectable orthologs in *Br* subgenomes I and III.

Of the 16,330 *At* genes used in *At–At* CNS discovery, we identified at least one *Br* ortholog for 6,245 *At* genes, with 2,391 *At* genes having a single detectable orthologous copy in *Br* (singlets), 1,723 *At* genes with two orthologous copies in *Br* (doublets) and 654 *At* genes with three orthologous copies in *Br* (triplets). We expected to find, in the absence of mutation, an αCNS whenever its gene was present. So, each gene in a

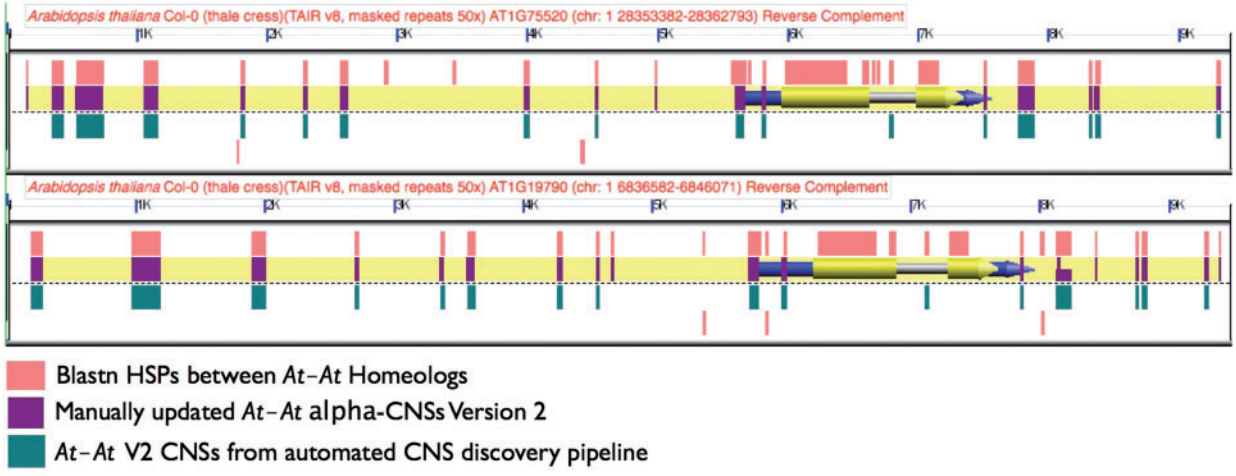

Fig. 2.—A CNS-rich pair of genes in *Arabidopsis*, retained from the most recent (α) paleotetraploidy, compared as sequence using "find CNS" blastn settings and displayed in the GEvo viewer. Panels of genomic regions (which can be regenerated at http://genomevolution.org/r/4db1, last accessed March 23, 2013) annotated using the GEvo application in the CoGe suit of tools (http://coge.iplantcollaborative.org, last accessed March 23, 2013). The figure compares an *At* gene (*AT1G75520*), a member of SHI transcription factor gene family and its homeolog. Blastn HSPs between the two genes (orange rectangles), manually updated αCNSs (purple blocks on upper model line; V2, supplementary table S1, Supplementary Material online) and CNSs detected using automated CNS pipeline (green blocks on lower model line; supplementary table S3, Supplementary Material online) are annotated in this figure. Note the similarity of the two CNS annotations, and how the HSP data in this experiment generally supports our CNS calls.

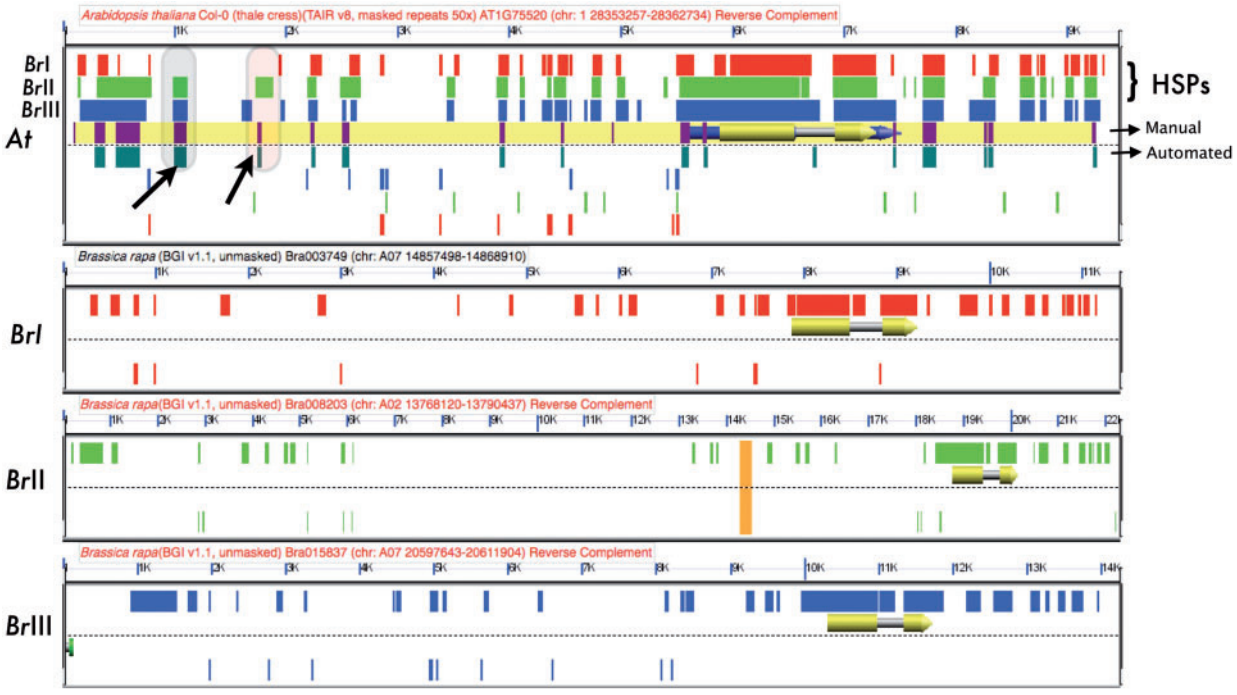

Fig. 3.—The fates of different *Arabidopsis* CNS sequences from figure 2 in the three subgenomes of *Brassica rapa (Br)* visualized in GEvo blastn comparison. Regenerate this experiment at http://genomevolution.org/r/4db6, last accessed March 23, 2013); the GEvo application (http://coge.iplantcollaborative.org, last accessed March 23, 2013). The top panel shows the αCNS-rich *At* gene (*AT1G75520*) of figure 2, a member of SHI gene family, and the three panels below show its detectable orthologous genespaces in *Br* (*Br* I, II, and III). αCNSs (purple bars) and the gene space (yellow background) are annotated on the *At* gene panel. HSPs corresponding to pairwise blastn comparison between the *At* gene and each of the three panels are indicated on the *At* gene panel as red bars for *Br* I, green bars for *Br* II and blue bars for *Br* III); the default color scheme in GEvo differs. Gray area highlighted follows the detectability of one αCNS across all three *Br* orthologous regions. Orthologous copies of this CNS are detectable in *Br* II and *Br* III subgenomes, but undetectable in *Br* I.

doublet or triplet would have an expected CNS. Based on *At–Br* orthologies, we expected to find 9,179 CNSs within the expected orthologous positions, 3,882 as singlets, 3,678 as doublets and 1,619 as triplets. Our detectability results are in supplementary table S2, Supplementary Material online. Many mutations to nondetectability occurred.

αCNS Length vs. Detectability

The version 2 CNS collection includes CNSs as short as 15 bp and as long as 283, and each has an *E* value more significant than that of a 15/15 exact nucleotide match. Even one base substitution would render some of these sequences undetectable using our blastn settings, so we expected that detectability would increase with length, and it did. αCNS length versus detectability was plotted for all version 2 αCNSs. Figure 4*A* shows these data for *Br* ortholog singlets, doublets, and triplets. In general, detectability is greater in singlets then doublets than triplets, as expected from our previous results and our general understanding of purifying selection and CNS redundancy. For singlets, detectability increases from 40% for 15–19 bp to 96% for more than 76 bp, with the 31–40 bp bin being 85% detectable. For the bin 51–75 bp, detectability was 62%, 72%, and 91% for triplets, doublets, and singlets, respectively. We chose those 2,509 αCNSs that are more than 30 bp in length to analyze further as to the molecular mechanism of their loss of detectability.

Pooling all αCNSs that are 31 bases or longer, we compared the degree of detectability in *Br* as a measure of the number of expected orthologous copies; we compared singlets with doublets with triplets. Figure 4*B*: Each αCNS is localized to one of the nine "categories" of *Br* genome: singlet subgenome I, singlet II, singlet III, doublet I, doublet II, doublet III, triplet I, triplet II, and triplet III. Figure 4*B* includes numbers of genes in each category, and probabilities that particularly interesting differences are significantly different. Detectability for CNSs on singlet genes is generally greater than that for doublet or triplet; that is expected because it should be more difficult to remove a singlet CNS without removing function. Detectability of CNSs on singlet genes of subgenome III is approximately 100%, and is significantly more than detectability of singlets on subgenomes II and I. There is certainly subgenome bias in the detectability of CNSs. This interesting result is not easy to explain, is probably important, and will be discussed.

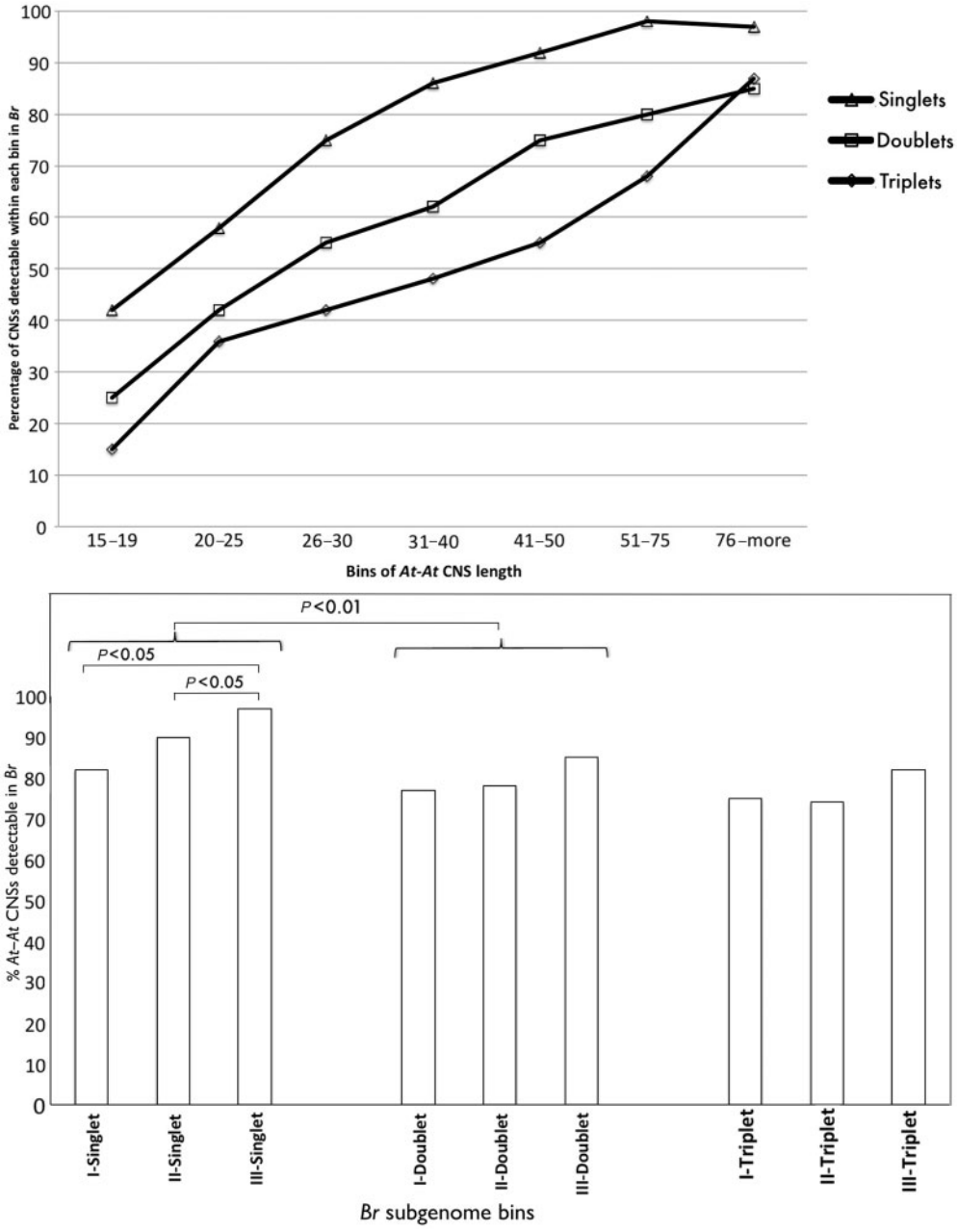

Fig. 4.—(*A*) CNS length versus detectability. Percent detectability of αCNSs in *Br* over bins containing CNSs of a given length (base pairs). Expected number of copies for each αCNS is based on number of detectable syntenous orthologs for each *At* gene in *Br* genome. (*B*) Comparison of detectability of all expected copies (singlet, doublet, or triplet) of *At–At* CNSs (31 bases or longer) between the three subgenomes within *Br*. Expected number of copies for each *At–At* CNS is based on number of detectable syntenous orthologs for each *At* gene in *Br* genome.

The Functionality Metric: Deletion to Loss-of-Function Is the Primary Mechanism for Removal of αCNSs More Than 30 bp Long in *Br*

Having located the stretch of chromosome in *Br* where the missing CNS could be, we devised a global alignment algorithm, global-npe, to identify the predominant mechanism of removal of CNSs. Table 1, column 1 lists these predominant mutational causes for the failure to detect a CNS. For each mechanism, impact on detectability within singlet genes was used as a control, and recorded as data in column *A* of table 1. We then sorted the 498 redundant (doublets and triplets), undetectable CNSs into mutational mechanism categories, and recorded these data in column *B*. *B*/*A* is the functionality metric, with a fully functional CNS category scoring 1, by

definition. *B/A* ranged from 5.3 (nonfunctional deletions) to 1 (fully functional). The functionality metric is useful. As deleted DNA cannot be functional, we now know that computed deletions have a 64% chance of being nonfunctional. Base substitutions, however, have a 91% chance of still being functional (but there are not many CNSs that have gone undetectable for this reason). Insertions may destroy detectability, while function is almost always maintained. Overall, the average nondetectable, once-redundant CNS has a 72% chance of being nonfunctional, and that makes sense because 97% of them sort to a molecular mechanism with "deletion" in its description. Deletion is the predominant mutational mechanism for the lack of detectability of αCNSs in *Br*, but other mechanisms operate as well.

Although our functionality metric differences imply that genes with undetectable CNSs generally function, it is more rigorous to test directly to see if the loss of CNSs is somehow correlated with the loss of gene function. Cheng et al. (2012) published RPKM values in leaves, stems, and roots of seedling *Br* Chiifu; our subgenomes I, II, and III are their subgenomes MF1, MF2, and LF. Using two different cutoffs for potential gene death, there was no correlation between loss of CNSs and potential gene death (supplementary table S3, Supplementary Material online). There was a slight tendency for subgenome I to have more dead genes than other subgenomes, and, as expected, the more stringent cutoff found fewer (~7%) potentially dead genes as compared more potentially dead genes (~17%) for the permissive threshold. Conclusions from this control experiment: CNS nondetectability or even total CNS loss is not correlated with gene death, so there is no need to modify the predictions of the functionality metric of table 1. Note that three organs in one environment do not monitor all of the possible expression endpoints, so the frequencies of genes that are actually dead are definitely below these potential death values (supplementary table S3, Supplementary Material online).

Detectability of CNS-Enriched TFBS Motifs, Especially the G-Box, in *Br* Orthologous Positions

TFBS motifs contained within *Arabidopsis* αCNSs, and enriched more than 2-fold within CNSs, were studied without regard to whether their CNSs were detectable in *Br*. They were detected as an exact match in the expected genespace region. Were any such motif in *Arabidopsis* lacking function in the *Brassica* lineage, base substitution alone (*At–Br* Ks = 0.38) would likely lead to nondetectability: a 5mer would become undetectable 91% of the time, and a 6mer, like the G box, would go undetectable 95% of the time if base substitution were the only mutational mechanism operating (which is certainly not the case). Supplementary table S4, Supplementary Material online, updates αCNS enrichment data from 2007 (Freeling et al. 2007)—using version 2 CNSs. For our TFBS detectability and enrichment studies, we included all 11,448 *At* αCNSs, not just those more than 30-bp long. We studied 12 motifs, and 8 of them contained the 5′ACGTG core (colored red in supplementary table S4, Supplementary Material online); this core is part of the G-box. Of these, each motif was counted as complement plus reverse-complement. The most enriched motif was the G-box (CACGTG palindrome) at 12.9-fold. 12.9 times more G-boxes are in αCNSs than in nontransposon, noncoding, non-CNS control space, normalized by position relative to the gene. The six base pair G-box derivatives plus core were all significantly enriched, but at values as low as 6-fold. The four not-G-box motifs were significantly enriched at between 2.4- and 8.1-fold. We found a poor correlation between motif enrichment and detectability in *Br* for these 12 motifs, although the G-box itself—most enriched—was third highest in detectability at 63%. Higher than the G-box in detectability was the 5′CCGTCC "meristem" box at 65%, with an enrichment of 8.1-fold. The jasmonic acid box, 5′GCCGCC, enriched to a paltry 2.4-fold, was relatively highly detectable at 50%, and highly enriched G-box-core 8-mer derivative 5′C**ACGTG**GC was detectable in *Br* only 26% of the time. Detectability is certainly giving us clues as to what sequence is essential for any generalized function, and what sequences may be superfluous, as will be discussed. For example, the 5′ACGTG G-box core was the most detectable motif of all, at 67.5%. In the absence of additional information, it seems obvious that some motifs may function in *Br* even though mutated while others have more absolute requirements for continuing function. The G-box itself is a CNS-enriched motif that seems to have a requirement for near-perfect sequence conservation to preserve function, and is especially intolerant to changes in the 5′ACGTG core. Therefore, the G box is a known motif that should be useful to study to independently determine the mutational mechanism that causes nondetectability.

Detectability Studies for G-Boxes in αCNSs, CNSs That Are Particularly Likely to Function

As we already provided evidence that 75% lost CNSs were deleted and thus, were mutated to no function, our premise is that G-boxes are primarily lost by deletion. Given our ability to pull-out and analyze expected orthologous regions within *Br* for comparison with the corresponding conserved noncoding space in *Arabidopsis*, we looked at all CNS-enclosed G-boxes and, more importantly, a subset of these that were experimentally shown to function in light regulation. PIL5 is an *Arabidopsis* transcription factor of the basic-helix-loop-helix type that is known to bind sequence containing a G-box. Oh et al. (2009) used microarray data and ChIP-chip (fragments from chromatin immunoprecipitation were localized by hybridization on microarrays) data to infer that, of the 748 *Arabidopsis* genomic binding sites occupied by PIL5, 166 genes were upregulated in light directly by PIL5. Each PIL5 site represents a strong argument for a functional

G-Box. We compared these PIL5 sites for overlap with our αCNS data set and identified 27 αCNSs containing a PIL5-informed G-Box. Each CNS was traced in all expected orthologous positions in *Br*. For each undetectable G-Box, we used our global-npe alignment to identify the predominant evolutionary mechanism that mutated the motif to undetectability in the *Br* lineage. Again, G-boxes were followed for detectability independent of whether the CNS expected to carry them was detected in *Br*. Figure 5 gives these results for all αCNS G-boxes and for the 27 G-boxes comprising the "most likely to function" subset, side-by-side. Our overall result: deletions—not point mutations, indels or insertions—removed the detectability of the majority of G-boxes: 73% for CNS-contained motifs and 82% for PIL5-informed G-boxes. Base substitutions account for a smaller but significant portion, approximately 15%, of G-box mutations to nondetectability.

Discussion

Purifying Selection in *Br* Resulted in Many Gene Regulatory Regions That Have Lost Cis-Acting Binding Sites, and 75% of the Time, These Sites Were Deleted and Therefore Have No Chance to Function in the Ancestral Manner

Table 1 summarizes the complete CNS detectability data of supplementary table S1, Supplementary Material online. Based on the length versus detectability data of Figure 4*A*, we demanded that CNSs be more than 30-bp long for our focal analysis of table 1. We reasoned that CNSs localized to genes that were fractionated down to one (singlets) would loss their genes rarely; from table 1, this "loss" frequency was 10%, and this became our least-redundant control pool. (This frequency of 10% is not negligible and is discussed in the next section.) Those CNSs that existed near doublet and triplet genes are expected to be relatively more redundant and more liable to loss-of-function mutation, so these CNSs became our experimental pool. We expected that more-redundant CNSs should go undetectable by whatever mutational mechanisms operated in the *Br* lineage to a greater frequency than they go undetectable in the singlet controls. This was indeed the case (fig. 4*A* and *B*; supplementary table S1, Supplementary Material online). Overall, an *At* αCNS more than 30-bp long, either 5′ or 3′ of its gene, mutates to undetectability in *Br* 33% of the time (table 1, last row, column *B*). The functionality metric for those CNSs that go undetectable by proved chromosomal deletions was 5.3, becoming our maximum not functional value; the CNS must be nonfunctional because the original DNA is not there. A functionality metric of 1 indicates complete functionality because redundancy makes no difference; nondetectability by large insertion had a negligible effect on functionality. The functionality matrix for those CNSs going undetectable because of base substitutions was 1.8, meaning that only 34% of CNSs in this category lost function, 66% of them still functioned even though they were undetectable. However, only 1% (16/1,543) more-redundant CNSs (column *B*) CNSs went undetectable for this reason. Considering all 498 cases where a more-redundant CNS went undetectable in the *Br* lineage (table 1, last row), 72% of these went nonfunctional, as expected because they were largely placed in categories characterized by the word "deletion."

There was no correlation between CNS loss and potential gene death (supplementary table S3, Supplementary Material online). Different branches on the plant phylogenetic tree have differed greatly in transposon blooms and polyploidies. Although there is no experimental evidence, it is possible that the rate of deletion and/or the size of the average deletion differs greatly among plant lineage, so extrapolating from our "mostly deletions" conclusion in *Brassica* to other plant lineages is not warranted. Interestingly, researchers in the detectability of ultra-CNSs in vertebrates noticed that post-paleotetraploid teleost fish lost CNS detectability much faster than sister vertebrate lineages not undergoing polyploidy (Lee et al. 2011). Again, an "induction" relationship is possible, but not proved.

The deletion mechanism we envision is the intrachromosomal recombination mechanism discovered for transposons in *Drosophila* (Petrov et al. 1996), described for transposons and genes, respectively, in maize (Devos et al. 2002; Woodhouse et al. 2010), evidenced in *Br* as rare exons carrying deletions (Tang et al. 2012), in rice (Tian et al. 2009) and inferred here to be the prevailing mutation mechanism in *Br*. The importance of short direct repeats flanking deleted DNA was first shown as a RecA-responsive process in bacteria (Albertini et al. 1982). Not all deletions need to be caused by the same mechanism. Some deletions may be mediated by flanking transposons and/or mis-repair of gaps caused in the movement process (Wicker et al. 2010). Similarly, strand slippage in the replication fork could generate short intrachromoaomal recombination deletions (Petrov 2002). Whatever the mechanism, the fact that we often see kilobase stretches of *Br* (and in maize: Woodhouse et al. 2010) removed when a gene and all of its CNS are fractionated does not mean that deletions in plants are long. It seems obvious that, once an initial deletion renders the gene functionless, then some combination of [rate of deletion] and/or [length of deletion] will incrementally remove the entire cis-acting unit.

There has been enough work in animal rates and lengths of deletion to permit a gross comparison of our *Brassica* lineage deletion process and that operating in any animal studied. There are no examples of an ordinary gene being lost in the human lineage by deletion; all are still present in situ as pseudogenes (Schrider et al. 2009); the human–chimpanzee is about as diverged as are the *Br* subgenomes! When genes were lost from the pheromone network in old word apes, the genes remain as obvious pseudogenes (Liman and Innan 2003); they were not deleted. Petrov (2002), in a theoretical

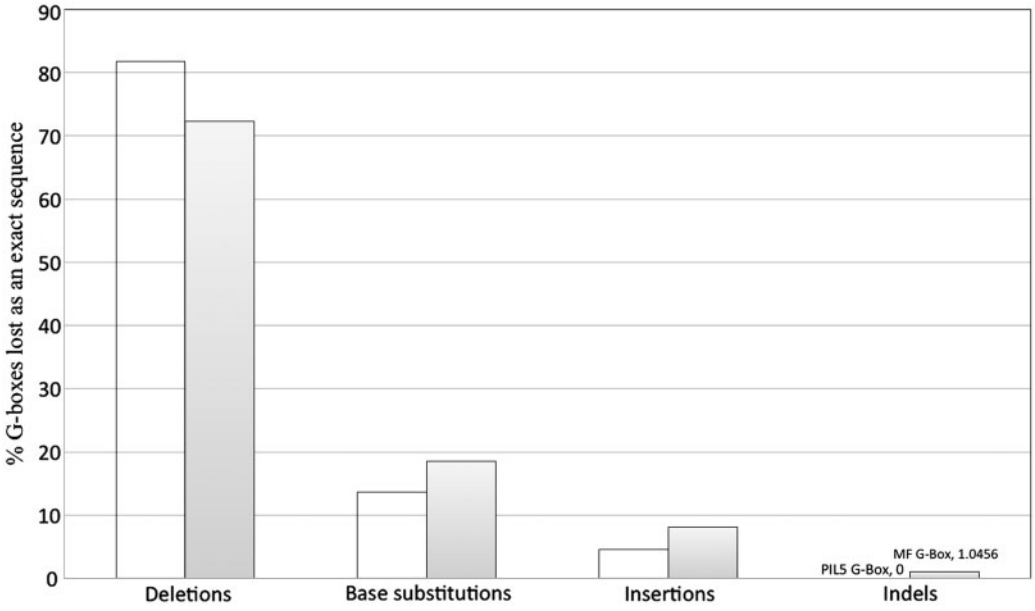

Fig. 5.—Predominant mechanism causing lack of detection of the G-Box (CACGTG) within αCNSs within the expected orthologous segments of *Brassica rapa*. Data for G-Boxes detected using a regular expression are labeled MF-G-Boxes; data for PIL-5 defined G-Boxes detected within αCNSs are labeled as PIL5-G-Box. Pil5-G-boxes are a subset.

essay on how the *C*-value paradox is best solved by a balance of deletion and insertion, reviewed the data of others on average rate of deletion per bp substitution and the average size of these deletions. He reviewed data in *Drosophila*, *C. elegans*, crickets, primates + rodents, and grasshoppers. Average rates of deletions per base pair substitution ranged from a low of 5% in the mammals to a high of 8.7% in *Drosophila*. The average size of a deletion was more variable, from a low of 1.6 bp in grasshoppers thru 3.2 base pairs for the mammals to a maximum of 48 bp for *Caenorhabditis elegans*.

For the purpose of illustration, imagine a 10-kb stretch of DNA that used to contain an entire gene and cis-acting elements. In the *Brassica* lineage, a Ks = 15% (the *Br–Br* Ks, fig. 1) is enough divergence time to remove the entire gene-space—exons plus all CNSs—without a trace; this occurred routinely during gene fractionation after the paleohexaploidy (Wang et al. 2011). Using Petrov et al.'s (1996) maximum animal deletion rate/bp (0.13 for *Drosophila*) and *Drosophila*'s deletion length of 38 bp, This 10 kb of functionless DNA would suffer $0.15 \times 0.13 = 2\%$ of its bps, or 200 deletion events, each averaging 35-bp long, giving 7 kb of deletion. Using the mammalian rate and length, and the earlier-mentioned crude arithmetic method, primates and rodents would delete only 240 bp of the 10 kb. The *Brassica* lineage uses a combination of deletion rate and length to more rapidly delete its functionless DNA than animals with tiny genomes and vast population sizes, and far more rapidly than do primates and rodents. The most obvious difference between plants and animals are the hundreds of millions of pollen shed per plant, each grain being a haploid gametophyte. Somehow, the plant's biology accommodates the "genetic load" commensurate with its relatively strong pressure of purifying selection.

Even Fully Fractionated *Br* Genes (Singlets) Are Likely to Be Functionally Redundant at Least 10% of the Time

All of the CNSs used in this study were from homeologous gene pairs retained from the most recent tetraploidy in the *Arabidopsis* lineage (modal Ks 76%, fig. 1, called alpha). Thus, each of these CNSs is redundant or nearly so in *Arabidopsis*. When this gene is triplicated as part of the paleohexaploidy in the *Br* lineage, there are originally three copies of the progenitor gene and its associated CNSs. If all but one of these genes fractionates, one might guess that the remaining one gene, and each of its CNSs, would confer some nonredundant, unique function. However, 10% of the CNSs expected to be with these singletons are undetectable. In fact, the assessment of "nonredundant" is probably not the whole truth. For every singleton *Br* gene used in this study, there is a possible out-paralog (Koonin 2005)—the descendants of the *Arabidopsis* alpha pair—gene family in *Br* that can be a singlet, doublet, or triplet. This out-paralog lineage is included in figure 1 as grayed-out. In some cases, it may not matter which of these genes is active as long as a "correct" contingent of them are retained to specify optimal product levels. As a general test of this reasoning, we analyzed supplementary table S2, Supplementary Material online, and asked, "do singleton genes in *Br* have more first cousin genes (decedents of

this gene's α-pair) with retained doublet and triplets?" In other words, did the genome compensate for the loss of an *Arabidopsis* gene by amplifying retention of the "out-paralog" swout-paralog gene, (Koonin 2005) such that we should really consider all six potential *Br* orthologs of the alpha pair when we are studying dosage relationships. The data (from supplementary table S2, Supplementary Material online): When an *At* gene is retained as a singlet ortholog in *Br*, there is at least one detectable out-paralog 87% of the time. However, if the *At* is retained as a doublet, then at-least-one out-paralog retention drops to 75%, and if retained as a triplet, the out-paralog retention frequency drops further to 72%. Clearly, there is some overlap of function among the six immediate descendants of an alpha pair. Consider further: Before alpha in the Brassicales phylogenetic tree, the beta duplication generated another potential of six "out-paralogs" that we have not yet included in our conceivably dose-sensitive networks (fig. 1). For these reasons, we are careful to note that a "singleton" gene in *Br* is not "nonredundant," but "less redundant".

The sort of reasoning used above is complicated because there is an overall expectation that—for gene functions requiring a fixed stoichiometry of product level—many genes will be selected for maintaining the status quo of product balance (Birchler and Veitia 2010). However, there have been multiple polyploidies in the lineage of all plants (fig. 1) that certainly led to functional redundancy. Add to this complexity the fact that genes on over-fractionated subgenomes are expected to—on average—express to lower RNA levels than do genes on the less fractionated homeolog (Schnable et al. 2011). For *Br*, subgenome III is the dominant subgenome (Cheng et al. 2012).

αCNSs Generally Confer Function

Plant CNS function is supported by conservation itself (Lockton and Gaut 2005; Reineke et al. 2011), the association of CNS-richness with particular genes and motifs (Freeling and Subramaniam 2009), the positive association of CNSs with open chromatin (Zhang et al. 2012) and by expression association studies (Spangler et al. 2011). This study approaches the function question by comparing less-redundant CNS loss—those on singlet *Br* genes—with the loss of more-redundant CNSs, as they are expected to exit on doublet and triplet *Br* genes. Figure 4*A* shows the relationship between detectability and CNS length: with one exceptional data point, CNSs expected to be on singlet genes are more detectable than CNSs expected on doublets are more detectable than on triplets. This makes sense if purifying selection is strongest when there is only one copy, moderately strong when there are two copies and weak for triplets. For the bin carrying the shortest CNSs, detectability is 2.9-fold higher for a singlet than for a triplet. For the bin carrying the median-lengthed CNS (31–30 bp), a singlet is 1.8-fold more detectable as a singlet than as a triplet, with the doublet in the middle. This result implicates selection—and αCNS function—unless mutation rates are somehow correlated with the redundancy of cis-acting regulatory units motifs; that is not reasonable.

The results of figure 4*B* involving redundancy versus detectability are, in general, expected, but the differences in detectability of singlet CNSs depending on the subgenome (I vs. II vs. III) is disturbing. That subgenome III always has more detectable CNSs than do the other two subgenomes cannot be ignored. As with the data of figure 4*A*, purifying selection seems to act most strongly on CNSs that are less redundant. Removing a unique CNS from a gene could well remove an essential or selectable function. However, why should it matter on what subgenome the singlet αCNS is located? It does. Singlets on I, II, and III are detectable at 82%, 87%, and 96% with differences significant, $P < 0.05$. Although none of these differences is much smaller than 100%, there still must be an explanation. The most obvious is that subgenome III is less mutagenetic; it deletes at a lower rate than the other subgenomes for some structural reason. This "mutationist" hypothesis was very much in contention as an explanation of biased fractionation, where one subgenome's genes gets deleted significantly more often than the other subgenome (Sankoff et al. 2010; Freeling et al. 2012). However, this mutationist alternative was considered carefully and disproved unequivocally in the case of the maize paleotetraploid (Schnable et al. 2011). It was shown that both subgenomes of maize suffer mutations—deletions via intrachromosomal recombination (Woodhouse et al. 2010)—at the same rate but that one subgenome expresses its genes to higher levels than the other on average, so pairs of genes tended to fractionate the homeolog that expresses least. These workers (Schnable et al. 2011) did not just demonstrate that genome dominance predicts biased fractionation; they actually tested the rate of deletion of functionless transposon and intron DNA between subgenomes and found the rates to be the same. For these reasons, the "selectionist, not mutationist" explanation was adopted for the maize lineage tetraploidy, and predicted to apply to the paleohexaploidy in the *Brassica* lineage as well. For *Br*: fractionation is biased with subgenome III being the least deleted (Wang et al. 2011), the mechanism of exon loss is deletion (Tang et al. 2012), and subgenome III dominates its RNA levels over subgenomes I and II (Cheng et al. 2012), just as is the case of maize. Perhaps singlets on subgenome III carry genes that are particularly and continuously important for growth and development, and singlets on the not-dominant subgenomes just do not matter quite as much.

The G-Box and Motif Detectability

Figure 5 graphs detectability in *Br* (as an exact motif sequence) of 1) G-boxes within all *At* αCNSs and 2) a subset of these G-boxes that are also experimentally validated (by ChIP-chip) PIL5 helix-loop-helix transcription factor binding sites. As expected from the CNS detectability results (table 1), G-boxes that lose

exact sequence are almost always deleted, not lost by base substitution.

Supplementary table S4, Supplementary Material online, presents our update of CNS enrichment values given our version 2 of the *At* αCNS list, and slightly updated methods. This table also presents detectability data for all motifs enriched significantly in CNSs by more than 2-fold. For those motifs, it is important to know that, in every case, complement and reverse complement were enriched to an equal degree (by chi square). The 5-mer within the G-box, 5′ACGTG is more detectable than the G-box itself, and is more detectable than any of the 11 G-box derivatives. For example, the most CNS-enriched G-box derivative, 5′CACGTGGC and its reverse complement, was among the least detectable at 26%; we conclude that this 8-mer motif contains alternative bp substitution sequences that still function even when mutated for the majority of similar sequences, and we draw a similar conclusion for most of the G-box derivative motifs. Although these particular G-like-boxes may (or may not) be the optimum DNA-binding partner for one or a few protein–DNA interactions, this sequence is really not a motif. Rather, they are each a specific sequence that contains a motif. We suggest that this G-box situation is typical of the generally overlapping, multiple motif data that comprise our current plant motif lists. For example, one such list that attempts to be exhaustive—the 426 regular expression motifs gathered together in the MotifView application in CoGe—fall into many sets of overlapping sequences, each supported by a unique experimental datum. Detectability measures over evolutionary time may help consolidate binding sequences into actual motifs. By this reasoning, the "G-box" is not an actual motif, but a derivative. The actual motif by this reasoning, the core shared by all or most related sites, could be 5′ACGTG because it is the most detectable of CNS-enriched boxes in the "G-box" family.

The general aim of bringing together CNSs and motifs or clusters of motifs—and especially ChIP-seq sites (much needed data)—is not even well formulated for plants. We know next to nothing about what proteins actually bind CNSs, how many different binding sites generally occupy CNSs, or if the spacing of sites within or among CNSs is important.

Fractionation: Nature's "Deletion Machine"

Knowing that the predominant reason *Arabidopsis* CNSs go undetectable is deletion, leading to loss-of-function, is crucial for the intelligent application of a new strategy for enhancer-like site analysis: fractionation mutagenesis. For example, the *Br* paleohexaploid informs intelligent, hypothesis-driven enhancer experiments in *Arabidopsis*. Fractionation mutagenesis is exemplified in the GEvo blastn output graphic of figure 3, where the query is *Arabidopsis At1G75520*, a particularly CNS-extensive gene encoding a RING zinc finger protein of SHI-type. Its 17 CNSs, covering 7.5 kb of chromosome in addition to the 1.9 kb transcriptional unit, have been largely retained in triplicate in *Br*. However, fractionation has rendered undetectable—probably deleted—a few longer orthologs of αCNSs: those circled in figure 3 are clearly present in *At* and *Br II*. The arrows indicate individual sequences in *At*. If there were a particular RNA-level pattern that was missing or aberrant in *Br* I and *Br* III, but ancestral in *Br* II, the CNS denoted by the rightmost arrow would become a candidate sequence with a hypothesis as to its meaning. Looking further to the right in figure 3, we find that CNSs on subgenome III (*Br III*) are ancestral, but some CNSs are missing from orthologs on subgenomes I and II. As subgenome III is the dominant subgenome (Cheng et al. 2012), this bias for loss is expected.

There will soon be many more *orthologous At* CNSs when usefully diverged *Brassicaceae* genomes are sequenced and aligned, and when CNSs obtained from multiple alignment data are merged with our pairwise CNS list. Those fractionated in *Br* should predominately lose function. It is valid to think of the fractionations following polyploidy in the Brassicas (and probably following other plant polyploidies as well) as deletion machines ideally suited to be used in the procedure of fractionation mutagenesis.

Supplementary Material

Supplementary tables S1–S4 and figure S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by National Science Foundation grants DBI 0701871 and IOS-1248106 to M.F. The authors thank Eric Lyons for helpful criticism and advice along the way.

Literature Cited

Albertini AM, Hofer M, Calos MP, Miller JH. 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29:319–328.

Baxter L, et al. 2012. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell 24:3949–3965.

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol. 186: 54–62.

Cheng F, et al. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One 7:1–9.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 19:1114–1121.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 127:1075–1079.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:433–453.

Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. 2007. G-boxes, bigfoot genes, and environmental response: characterization of

intragenomic conserved noncoding sequences in *Arabidopsis*. Plant Cell 19:1441–1457.
Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol. 12:126–132.
Freeling M, et al. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol. 15:131–139.
Hancock JM, Shaw PJ, Bonneton F, Dover GA. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol Biol Evol. 16:253–265.
Hollister JD, Ross-Ibarra J, Gaut BS. 2010. Indel-associated mutation rate varies with mating system in flowering plants. Mol Biol Evol. 27: 409–416.
Junion G, et al. 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148:473–486.
Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. Proc Natl Acad Sci U S A. 99:6147–6151.
Koonin EV. 2005. Orthologs, paralogs and evolutionary genomics. Annu Rev Genet. 39:309–338.
Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol. 283:1205–1215.
Liman ER, Innan H. 2003. Relaxed selective pressure on an essential component of pheromone transduction in primate evolution. Proc Natl Acad Sci U S A. 100:3328–3332.
Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 21:60–65.
Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.
Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53: 661–673.
Moses AM, et al. 2006. Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol. 2:1219–1231.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48:443–453.
Oh E, et al. 2009. Genome-wide analysis of genes targeted by phytochrome interacting factor 3-like5 during seed germination in *Arabidopsis*. Plant Cell 21:403–419.
Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko A. 2007. Predicting tissue-specific enhancers in the human genome. Genome Res. 17: 201–211.
Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature 384:346–349.
Petrov DA. 2002. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61:531–544.
Raatz B, et al. 2011. Specific expression of lateral suppressor is controlled by an evolutionarily conserved 3′ enhancer. Plant J. 68:400–412.
Reineke AR, Bornberg-Bauer E, Gu J. 2011. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res. 39:6029–6043.
Sankoff D, Zhen C, Zhu Q. 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11:313–324.
Schnable JC, Pedersen BS, Subramaniam S, Freeling M. 2011. Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses. Front Plant Genet Genomics. 2:1–7.
Schrider DR, Costello JC, Hahn MW. 2009. All human-specific gene losses are present in the genome as pseudogenes. J Comput Biol. 16: 1419–1427.
Spangler JB, Subramaniam S, Freeling M, Feltus FA. 2011. Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. New Phytol. 193:241–252.
Tang H, Lyons E, Pedersen B, Paterson A, Freeling M. 2011. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12:102–113.
Tang H, et al. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190:1563–1574.
Thomas BC, Rapaka L, Lyons E, Pedersen B, Freeling M. 2007. *Arabidopsis* intragenomic conserved noncoding sequence. Proc Natl Acad Sci U S A. 104:3348–3353.
Tian Z, et al. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. 19:2221–2230.
Van de Peer Y. 2011. A mystery unveiled. Genome Biol. 12:113.
von Rohr P, Friberg MT, Kadarmideen HN. 2007. Prediction of transcription factor binding sites using genetical genomics methods. J Bioinform Comput Biol. 5:773–793.
Wang X, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 43:1035–1039.
Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. 20: 1229–1237.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2:333–341.
Woodhouse M, et al. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 8:e1000409.
Zhang W, et al. 2012. High-resolution mapping of open chromatin in the rice genome. Genome Res. 22:151–162.

Associate editor: Bill Martin